基于CAT的在线标定：设计与方法

张雪琴[1]　毛秀珍[1]　李佳[1]

（四川师范大学教育科学学院，四川成都，610066）

摘　要　项目增补是题库建设和维护的重要手段，而标定新题参数是项目增补的重要内容。在线标定设计和在线标定方法分别研究新题的施测方式和参数估计方法，是计算机化自适应测验（computerized adaptive testing, CAT）情景下项目增补的核心技术。重点厘清在线标定设计与在线标定方法的发展思路和脉络，并对它们的特点、联系和表现进行介绍和评价。未来应基于其他信息指标进一步研究在线标定设计，可基于联合估计和误差校正的思路探究在线标定方法，应加强研究认知诊断 CAT 和多维 CAT 的在线标定技术，深入开展项目增补方法的实证研究。

关键词　计算机化自适应测验；认知诊断理论；项目增补；在线标定设计；在线标定方法

分类号　B841

1 引言

随着现代测量学理论和信息技术的不断发展，计算机化自适应测验（computerized adaptive testing, CAT）已成为心理与教育测量领域的重要分支，受到研究者们的青睐。CAT 根据被试当前的能力估计水平为被试选择最适合的题目作答，实现测验的“量体裁衣”、“因人施测”的自适应思想，从而在保证测量精度的前提下大大减少了测验的长度。除此之外，CAT 使得项目呈现更加标准化，实现实时评分，还能提高测验的安全性和公平性。

CAT 由事先完成标定的题库、初始项目选择、选题策略、能力估计方法以及终止规则五个部分组成(陈平等, 2013)。其中，题库是 CAT 的基础，其质量的高低将影响测验的安全性和准确性。当题库中的试题被长期使用时，一部分题目必定会因过度曝光、内容陈旧或本身有缺陷等问题，将不再适合继续使用(Wainer & Mislevy, 1990)。因此，有必要定期开发新题对存在问题的试题进行替换，并且新题必须经过传统纸笔试测或 CAT 的方式标定参数后才能纳入正式题库。特别地，在 CAT 测试过程中收集信息并估计项目参数，称为在线标定。陈平等人(2013)指出在线标定中考生同时作答旧题和新题，依据旧题参数估计考生能力进而估计新题参数，相当于锚人设计，这样就很自然地将新题参数置于旧题同一量尺上。因此，在线标定不再需要复杂的等值方法。与传统方法相比，在线标定因不需要组织单独的试测，

收稿日期：2020-04-19
通讯作者：毛秀珍，E-mail：maomao_wanli@163.com

大大节省了题库维护的成本和时间，减轻了测验开发者的负担，在大规模题库维护中具有重要的实践意义。

具体而言，在线标定是指在 CAT 测试中将新题嵌入测验的不同位置，当考生到达嵌入位置时，将新题以一定方式分配给考生作答，并收集反应数据用于估计新题参数的过程(Wainer & Mislevy, 1990)。在线标定包括新题的嵌入位置、选题方法(称为在线标定设计)、参数估计(称为在线标定方法)和终止规则四个方面（Zheng, 2014）。迄今为止，研究集中于新题的选题设计和参数估计算法，对新题的嵌入位置和终止规则的研究还比较薄弱。

首先，在线标定设计分为随机标定设计和自适应标定设计。早期研究者将自适应标定设计看作取样设计(Berger, 1992; Jones & Jin, 1994)，自适应的选取参与标定的最优考生样本。即根据项目特征“选人”的方法，包括 D-优化和序贯 D-优化(Berger, 1994; Chang & Lu, 2010; Jones & Jin, 1994)。现阶段，研究从“选人”过渡到“选题”，旨在为当前考生施测最适合其标定的项目。一方面基于项目信息标准建构了两点 D-优化设计(Ren et al., 2017)、贝叶斯 D-优化设计(van der Linden & Ren, 2015)、优秀度指标(He et al., 2019)和 D-c 设计(He & Chen, 2020)，另一方面基于考生样本约束提出适合度指标法(Ali & Chang, 2014)和区间序列信息优先指标法(Zheng, 2014)。

其次，在线标定方法在传统 CAT、多维 CAT(multidimensional CAT, MCAT)和认知诊断 CAT(cognitive diagnostic CAT, CD-CAT)都有研究。传统 CAT 中的在线标定方法主要分为条件极大似然估计(conditional maximum likelihood estimation, CMLE)(陈平, 2016; 游晓锋等, 2010; He et al., 2017; Stocking, 1988)、边际极大似然估计/期望极大算法(marginal maximum likelihood estimation via expectation maximization method, MMLE/EM)(Ban et al., 2001; Chen & Xin, 2013; Wainer & Mislevy, 1990)和贝叶斯估计方法(Chen, 2017; Zheng, 2014)。MCAT 主要以推广传统 CAT 的方法为主。CD-CAT 中包括项目参数估计(陈平, 辛涛, 2011a)、Q 矩阵估计(汪文义等, 2011)、联合估计项目参数和 Q 矩阵(陈平, 辛涛, 2011b; 谭青蓉, 2019; Chen et al., 2015)。

本文的第二、三部分重点对在线标定设计与在线标定方法的特点、表现和联系进行介绍和评价，厘清相关研究的发展思路和脉络；第四部分简单回顾新题嵌入位置和终止规则的研究进展。在此基础上，第五部分针对传统 CAT、CD-CAT、MCAT 在线标定设计、在线标定方法的理论与实践提出一些具体的研究方向和展望。

为了行文方便，先对文中的符号进行说明。首先，j 和 β_j 分别表示待标定的新题 j 及其参数向量；R_{new} 表示新题库，k-1 表示当前已作答新题 j 的被试人数，x_i^j 表示考生 i 在新题 j 上的作答反应，$\hat{\theta}_i$ 表示考生 i 的能力估计值。

2 在线标定设计

在线标定设计即新题的施测方式。如何将被试与新题合理搭配以优化题目参数标定的效率是在线标定设计的核心问题。Wainer 和 Mislevy (1990)最早提出随机和自适应两种在线标定的设计方式。其中，随机标定设计为每个被试从新题库随机选取固定数量的新题，并随机嵌入测验进行施测。游晓锋等人(2010)、汪文义等人(2011)和陈平 (2016)的研究都运用了随机标定设计。游晓锋等人(2010)研究发现随机设计中新题的作答次数越多参数估计越准确。随机标定设计简便易行，但嵌入新题的难度与相邻题目的难度可能存在明显差异，考生易察觉，造成不认真作答从而影响参数估计的精度。更重要的是，随机设计没有体现 CAT 自适应的优点。

自适应标定设计按标准选取最能反应项目特征的被试，或者选取最适合当前被试标定的新题施测，成为在线标定设计研究的新方向。根据在线标定设计指标建构的思路，可将在线标定设计分为基于项目信息标准和基于考生样本约束的最优设计两类。

2.1 基于项目信息标准的最优设计

利用项目参数的信息量来反映参数估计误差，是基于项目信息标准最优设计的基本思路。基于项目信息标准的最优设计主要包括 D-优化，序贯 D-优化，两点 D-优化，贝叶斯 D-优化，优秀度指标和 D-c 设计等。

2.1.1 D-优化设计方法及其改进

D-优化设计通过最大化项目参数 Fisher 信息矩阵的行列式来最小化项目参数的广义协方差，是一种以优化项目参数估计效率为目标的统计指标（Zheng, 2014）。当新题 j 已经被 k-1 个考生作答，选取第 k 个考生时，将选取使式（1）最大化的能力为 θ_k 的最优考生作答该项目，换言之，

$$\theta_k = \arg\max_{\theta}\{\det[\sum_{i=1}^{k-1} I_j(\beta_j, \hat{\theta}_i) + I_{\mathrm{j}}(\beta_j, \theta)], \theta \in \Theta\}, \quad (1)$$

其中，Θ 代表所有考生能力 θ 的集合，$I_{\mathrm{j}}(\beta_j, \hat{\theta}_i)$ 代表被试 i 提供给新题 j 参数向量的 Fisher 信息量，$\sum_{i=1}^{k-1} I_{\mathrm{j}}(\beta_j, \hat{\theta}_i)$ 代表前 k-1 个考生在新题 j 参数向量上累计的信息量。在 D-优化基础

上，Berger (1992)指出 2PLM 中使式(1)取得最大值的θ_j为$b_j \pm 1.542/a_j$。因此，选取能力估计值与θ_j最接近的被试施测新题j，称为两点 D-优化设计法(Two-point D-Optimal Design，D-Tp)。Chang 和 Lu (2010)基于最优能力准则，在不定长 CAT 中直接应用两点 D-优化设计法，并按序贯的方式选取被试作答新题，称其为序贯 D-优化设计。

D-优化方法虽体现了 CAT 自适应的特点，但它假设存在一个由所有考生组成的“静态考生库”，库中考生的能力已知，考生可被任意选用为题目标定的样本，忽视了真实 CAT 情景下，在某时间点参与测验的考生人数不可控，导致“考生库”难以建构，同时很难找到与θ_k准确匹配的考生。

基于 D-优化设计的不足，研究者一方面从单点能力优化设计扩展到能力区间的优化设计。例如，Hassan 和 Miller (2019)基于限制性最优设计的思想，提出按照在最优能力区间而不是最优设计点进行取样，称其为局部限制性最优设计。实验结果表明，限制性 D-优化设计比随机设计取样效率更高。另一方面，从“依题选人”的思想过渡到“依人选题”。例如，Ren 等人(2017)认为从单点能力抽样被试不能产生稳定的参数估计值，同时基于 D-优化和 A-优化（Buyske, 1998）视角将最优能力替换为最优作答概率，提出 D-Tp1、D-Tp2 和 D-Tp3 方法。具体而言，已知考生p在当前新题嵌入位置时的能力估计值为$\hat{\theta}_p$，那么 D-Tp1、D-Tp2 和 D-Tp3 三种方法分别选择使$\{|\hat{\theta}_p - \theta_j^{(1)}|^{\delta_{pj}}, |\hat{\theta}_p - \theta_j^{(2)}|^{1-\delta_{pj}}\}$、$\{|P_{pj} - 0.176|^{\delta_{pj}}, |P_{pj} - 0.824|^{1-\delta_{pj}}\}$和$\{|P_{pj} - 0.25|^{\delta_{pj}}, |P_{pj} - 0.75|^{1-\delta_{pj}}\}$最小的项目给当前考生作答。其中，当项目$j$作答奇数次时，令$\delta_{pj} = 0$，反之为 1。结果表明，这三种方法的参数估计精度无明显差异，D-Tp2 的正确作答概率相对 D-Tp3 更极端，容易造成所选试题难度与前后题目难度差异较大，不利于新题参数的估计(Ren et al., 2017)。此外，Kang 等人 (2020)还从丰富被试信息的视角，利用被试的反应时信息，在联合反应和反应时模型下考察 D-优化、A-优化和随机方法的表现。结果表明，在参数标定过程中增加反应时能明显提高参数估计的准确性和效率。

2.1.2 贝叶斯 D 优化设计

van der Linden 和 Ren (2015)在 D-优化基础上根据当前考生p的能力估计值$\hat{\theta}_p$，选择使当前累加考生样本与之前累加考生样本相比，能提供项目参数 Fisher 信息增量最多的项目给考生p作答，提出贝叶斯版本的 D-优化设计，称为 D-VR 设计。即，

$$j = \arg\max_{j \in R_{new}} \{\det[\sum_{i=1}^{k-1} I_j(\beta_j, \hat{\theta}_i) + I_j(\beta_j, \hat{\theta}_p)] - \det[\sum_{i=1}^{k-1} I_j(\beta_j, \hat{\theta}_i)]\} \quad (2)$$

其中，$I_j(\beta_j, \hat{\theta}_i)$ 在计算中是结合了被试能力和新题参数的先验信息。这种方法虽从“选人”过渡到“选题”，考虑了现实的可行性，并在大样本条件下的标定效率较高，但是 D-VR 更倾向于选择具有统计优势的项目而忽略了其它项目，导致作答每个项目的样本量不均衡。另外，这种设计也容易造成所选试题的难度与前后题难度差异明显。

2.1.3 优秀度指标和 D-c 设计

鉴于 D-VR 设计中到达嵌入位置的当前考生不一定是最优考生样本，He 等人(2019)结合 D-优化的思路对 D-VR 改进，提出以式（2）表示的最优考生在新题 j 上提供的信息增量为基准，衡量当前考生 $\hat{\theta}_p$ 相对于最优考生 θ_k 在标定新题 j 上的优秀程度，选取优秀程度最高的题目给当前考生作答，称为优秀度指标（excellence degree, ED）方法。即，

$$j = \arg\max_{j \in R_{new}} \frac{\det[\sum_{i=1}^{k-1} I_j(\beta_j, \hat{\theta}_i) + I_j(\beta_j, \hat{\theta}_p)] - \det[\sum_{i=1}^{k-1} I_j(\beta_j, \hat{\theta}_i)]}{\det[\sum_{i=1}^{k-1} I_j(\beta_j, \hat{\theta}_i) + I_j(\beta_j, \theta_k)] - \det[\sum_{i=1}^{k-1} I_j(\beta_j, \hat{\theta}_i)]}\} \quad (3)$$

实验表明，在所有条件下，ED 设计在参数估计精度和标定效率上都优于 D-VR 设计。He 和 Chen (2020)还从项目参数估计误差的角度提出了 D-c 设计。该方法根据当前考生 p 的能力估计水平 $\hat{\theta}_p$，从新题库中选取能使项目参数估计误差产生最大减少量的项目 j 施测给考生 p，即最大化（4）式，

$$j = \arg\max_{j \in R_{new}} \{\det[(\sum_{i=1}^{k-1} I_j(\beta_j, \hat{\theta}_i))^{-1}] - \det[(\sum_{i=1}^{k-1} I_j(\beta_j, \hat{\theta}_i) + I_j(\beta_j, \hat{\theta}_p))^{-1}]\} \quad (4)$$

其中，$(\sum_{i=1}^{k-1} I_j(\beta_j, \hat{\theta}_i))^{-1}$ 代表 k-1 个考生在新题 j 参数向量上累计的信息矩阵的逆。D-c 设计与 D-VR 和 D-Tp 相比，能产生更高的参数估计精度和效率。特别地，当加入新题参数先验信息时，贝叶斯版本的 D-c 设计更能提高参数估计的精度和效率（He & Chen, 2020）。

2.2 基于考生样本约束的优化设计

针对 D-VR 方法中参与标定的考生样本量不均衡的问题，Ali 和 Chang (2014)、Zheng (2014) 分别基于考生样本量和考生能力范围的角度对参与标定的考生样本进行约束，分别提出适合度指标(suitability index,SI)和区间排序信息优先指标(ordered informative range priority index, OIRPI)。

2.2.1 适合度指标

SI 方法将考生能力值划为 r 个区间，设每个区间权重为 ω_r，并约束每个能力区间中作

答新题 j 的目标样本量为 T_{jr}。该方法通过 T_{jr} 和实际取得的样本量 t_{jr} 去平衡每道试题在不同能力区间的样本量，即，

$$SI_j = [1/(\hat{b}_j - \hat{\theta}_p)] \cdot \prod_{r=1}^{R} w_r[(T_{jr} - t_{jr})/t_{jr}] \tag{5}$$

与基于信息量的 D-优化统计指标有所不同，SI 指标通过平衡相邻题目间的难度和不同能力区间的考生样本量来优化项目参数估计效率。研究表明，SI 指标能明显提高随机方法的标定精度，但 SI 方法的目标样本和权重都是主观设定，难以保证其代表性和科学性。

2.2.2 区间排序信息优先指标

Zheng (2014)提出的 OIRPI 方法首先将考生能力划分为 r 个区间，并令每个区间 r 的中间值为 θ_r。然后计算每个新题 j 在所有能力区间 r 的 D-优化信息值 $D_{jr} = \sum_{i=1}^{k-1} I_j(\hat{\theta}_i) + I_j(\theta_r)$，在此基础上将每个区间的 D_{jr} 标准化得到 S_{jr}。最后选取当前考生所在区间 S_{jr} 值最大的项目 j 施测给当前考生。即，

$$S_{jr} = \{\frac{D_{jr} - \min_{r \in R}(D_{jr})}{\max_{r \in R}(D_{jr}) - \min_{r \in R}(D_{jr})}\} \tag{6}$$

OIRPI 指标将新题在不同能力区间的 D-优化指标值标准化，并用于表示题目对不同能力区间考生的需求度，再为当前区间的考生选取需求度最大而不是基于信息量的指标值最大的题目施测。因此，OIRPI 考虑了其他题目对当前考生的需求度，解决了 D-VR 设计中因项目统计优势而导致的样本不均衡的问题。Zheng 和 Chang (2017)指出 OIRPI 提高了 D-优化、D-VR 和随机方法的标定效率。

注意到，自适应标定设计的指标都依赖于新题参数的初始值。针对新题参数的初始值，Wainer 和 Mislevy (1990)建议出题者依据主观判断给出新题参数的初始值；Markansky (2009)和 Zhu (2006)则提出两阶段设计来获取参数的初始估计值。于是，陈平和辛涛(2011a)在 CD-CAT 情景下运用两阶段法，将在线标定分为预标定和重新标定两个阶段，并比较了两个阶段的样本比例对标定结果的影响。

2.3 在线标定设计简评

在线标定设计指标反映了被试与新题的匹配度，其建构思路可以从两个角度分析（见表 1）。一方面，基于项目视角则是利用项目特征的信息量反映参数估计误差的思路建构在线标定设计指标，指标的建构从“选人”过渡到了“选题”。总体上，随着样本量和参数初始估计精度的提高，D-c 设计表现最好，其次是 D-Tp、D-VR 设计。另一方面，基于考生视角的

思路则是约束考生样本大小和能力范围来提高参与标定的样本的质量进而建构在线标定选题指标。思路不同，在参数估计精度和效率上的表现也就不同，随着嵌入位置不断向后，OIRPI 的表现优于 D-优化、D-VR 和随机设计（Zheng,2014）。特别地，自适应标定设计依赖于项目参数初始值，在 2PLM 和 3PLM 中，区分度参数为 0.5 时，随机方法和 D-优化估计精度最高，然后是 OIRPI，D-VR（Zheng & Chang, 2017）。因此，现有的研究虽多，但它们都容易受到样本大小、参数估计初始值、新题嵌入位置等因素的影响，由此造成结论存在不一致的情况，有待进一步的研究。

目前，在线标定设计集中于传统 CAT，而 CD-CAT 在线标定设计的研究很少。D-优化设计是基于信息标准的指标，最初是用于表征项目对被试提供信息量的一种选题策略，将其转换为被试对项目提供信息量指标并用于构建在线标定设计指标，给在线标定设计提供一种新的研究视角，必将成为今后研究在线标定设计的重要思路。因此，未来研究可以借鉴 D-优化设计的思想，考察将其它基于信息量选题指标转换到建构被试对项目的信息量指标的可行性，并考察其在不同测验条件下的表现；针对 CD-CAT 项目参数和 Q 矩阵的标定需求，借鉴传统 CAT 在线标定思路探索 CD-CAT 联合标定项目参数和 Q 矩阵的在线标定设计；深入探究样本量、能力估计精度、项目参数初始值以及新题嵌入位置对自适应在线标定设计的影响，为实践应用提供方法和借鉴。

表 1　已有的 CAT 中的在线标定设计

分类标准	方法	特点
项目视角：参数信息量	D-优化、序贯 D-优化	自适应选取被试
	D-TP、D-VR、ED 和 D-c 方法	自适应选取项目
考生视角：能力与样本量	OIRPI、SI 指标	

3 在线标定方法

在线标定方法即新题的参数估计方法。目前，在线标定方法针对传统 CAT、MCAT 和 CD-CAT 都有一定的研究。

3.1 传统 CAT 的在线标定方法

3.1.1 CMLE 方法

Stocking(1988)最初提出的方法 A(Method A)运用 CMLE 的思想，将能力估计值当做真值来标定新题参数。这种方法容易将能力的估计误差传递到新题的标定过程，导致新题参数产生偏差，出现参数量尺漂移，从而降低参数估计精度。于是，Stocking(1988)又在 Method

A 的基础上提出了方法 B(Method B)，并在测验中加入一部分参数已经标定好且与旧题在同一量尺上的锚题，再运用等值技术将新题参数置于旧题相同的量尺上。方法 B 解决了方法 A 中参数量尺漂移的问题，但增加了测验长度和等值计算，需要花费更多的时间和精力。

另外，陈平 (2016)和 He 等人(2017)分别基于不同方法校正方法 A 中能力估计误差。具体而言，前者分别运用“全功能极大似然估计”和“利用充分性结果估计”(Stefanski & Carrol, 1985)方法与 Method A 结合用于估计项目参数，得到 FFMLE-A 和 ECSE-A 方法。后者提出了一种改进的 Lord 偏差校正法，并与方法 A 结合，得到 MLE-LBCI- A 方法。研究表明，FFMLE-A、ECSE-A 和 MLE-LBCI- A 方法都能有效提高方法 A 的标定精度（陈平, 2016；He et al., 2017）。尤其是在短测验中，FFMLE-A、ECSE-A 与最优的 MEM 算法接近（陈平 2016）。此外，Chen 和 Wang(2015)还将 FFMLE 方法的思路应用到 MCAT 中，并与 M-Method A 结合，得到 FFMLE-M-Method A，并指出在所有条件下，FFMLE-M-Method A 方法的参数估计精度明显高于 M-Method A。

除上述方法外，游晓锋等人(2010)提出的单参数、双参数以及多重迭代 MLE 方法也是 CMLE 思想的直接应用。

3.1.2 MMLE/EM 方法

Wainer 和 Mislevy (1990) 基于 MMLE/EM 算法衍生出单循环 EM 算法(one-cycle expectation-maximization method, OEM)用于项目参数在线标定。OEM 方法包含了一个 E 步和 M 步。其中，E 步基于被试在旧题上的作答反应计算能力后验分布，M 步基于被试在新题上的作答反应和能力后验分布估计新题参数。OEM 通过两步实现新题参数估计，方法简单，但在参数估计过程中并未利用新题参数信息。

Ban 等人(2001)提出了多循环 EM 算法(multipie-cycle expectation-maximization method, MEM)解决迭代不收敛的问题。MEM 包含多个 OEM 循环，从第二个循环开始，同时利用考生在新、旧题的作答反应和新题参数的临时估计值来更新能力的后验分布，当前后两次项目参数估计值之间的平均绝对偏差小于预定精度就达到收敛，并结束估计。Ban 等人(2001)指出 MEM 参数估计的精度最高，其次是 OEM、Method B 和 Method A，但 MEM 的迭代周期可能较长，比较耗时。

基于边际极大似然方法，Chen 和 Xin (2013)将 OEM、MEM 和 Method A 推广至 MCAT；Kang 等人(2020)针对联合反应和反应时模型提出了边际极大似然（marginal maximum likelihood estimation,MMLE）和边际极大后验概率（marginal maximum a posteriori estimation, MMAP）方法。

3.1.3 贝叶斯方法

在线标定初期考生样本较少，EM 算法中参数估计不易收敛，为了缓解这一问题，Zheng (2014)在方法 A、OEM 和 MEM 方法中加入新题参数的贝叶斯先验信息，提出了 Bayesian-A、Bayesian-OEM 和 Bayesian-MEM 三种方法，并在三种单维 IRT 模型下比较了参数估计精度。结果表明，加入贝叶斯先验信息的三种方法均表现较好。其中，Bayesian-MEM 表现最好，它不仅能彻底解决参数不收敛的问题，还能提高参数估计精度，但迭代过程比较耗时。特别地，选取正确、合理的项目参数先验信息尤为重要。由此，Zheng (2014)建议采用旧题参数的先验分布作为新题参数的先验分布。Chen (2017)又将 Bayesian-OEM 和 Bayesian-MEM 贝叶方法用在 MCAT 中，得到 M-OEM-BME 和 M-MEM-BME 两种贝叶斯方法，并比较了多种在线标定方法，获得与 Zheng (2014)一致的结论，即加入新题参数先验信息能够明显提高参数标定的准确性和效率。

研究者还探究了多级评分项目的在线标定方法。例如，熊建华等人(2018)改进了传统 CAT 中的夹逼平均法和 MEM 方法并将它们推广至等级反应模型（graded response model，GRM）。Zheng (2016)和 Xiong 等人(2020)分别将 OEM 和 MEM 拓广到分步评分模型（generalizaed partial credit model, GPCM）和 GRM 模型。实验结果表明，在两个模型下，OEM 和 MEM 均表现出较好的估计精度。

3.2 CD-CAT 在线标定方法

与传统 CAT 不同，CD-CAT 除了项目参数外还需要估计 Q 矩阵。针对项目参数估计，陈平和辛涛 (2011a)将 MethodA、OEM、MEM 推广到 CD-CAT，并指出 CD-MethodA 最简单且标定精度最高。针对 Q 矩阵的估计，汪文义等人(2011)在新题参数已知条件下提出了 MLE、MMLE、交差法标定项目属性向量。其中，交差法利用集合的交运算和差运算夹逼出新题的 Q 矩阵，对知识状态估计精度要求极高。

针对项目参数和 Q 矩阵的联合估计，陈平和辛涛 (2011b)首次基于 IRT 中联合极大似然估计的思路，提出一种联合估计算法(joint estimation algorithm, JEA)。JEA 方法的第一步，给定新题的 q 向量和参数的初始值，采用 MLE（汪文义等, 2011）估计新题的属性向量 q；第二步，视上一步估计的 q 向量为真值，采用 CD-MethodA（陈平, 辛涛, 2011a）估计新题的项目参数；循环一，二步直到满足预先设定的收敛标准或最大循环数。该方法允许逐个标定新题，在大样本且项目质量较高时，表现出较高的估计精度。

接下来，Chen 等人(2015)在 JEA 方法的基础上提出了单个项目估计法（single-item estimation, SIE）。具体来说，采用 EM 方法为新题 j 计算在每一种可能的 q 向量下的项目参

数，再将项目参数看做已知，采用 MLE 找到最大似然值对应的 q 向量和参数即为该新题的 q 向量和项目参数的估计值。随后他们又在 SIE 的基础上提出同时估计多个项目的 SimIE 方法。结果表明，在 Q 矩阵和项目参数估计精度方面，SIE 和 SimIE 方法优于 JEA 方法。然后，谭青荣 (2019)提出了适用于多种诊断模型的广义在线标定方法，分别在 SIE 和 JEA 方法基础上基于项目先验信息提出 SIE-R 和 JEA-R 方法，并引入模型复杂性指标提出 SIE-R-BIC 和 JEA-R-BIC 方法，还基于作答分布间一致性的思想提出了 RMSEA-N 方法。谭青荣 (2019)指出，在 Q 矩阵和项目参数估计的精度方面，新提出的方法都优于已有的方法。

3.3 在线标定方法的简评

由表 2 可知，现阶段在线标定方法的研究集中于传统 CAT，并以单维二级评分的 IRT 模型的为主，在单维多级评分 IRT 模型下的研究较少，未来研究有必要在多级评分模型下比较各种方法的表现。针对 MCAT 和 CD-CAT 主要是将传统 CAT 的在线标定方法进行推广，尚未出现基于 MCAT 和 CD-CAT 自身结构特点的在线标定方法的研究。注意到 CD-CAT 中要么假设 Q 矩阵已知时估计项目参数，要么假设 Q 矩阵未知时联合估计 Q 矩阵和项目参数。于是，未来研究既可以基于校正能力估计误差的思路校正 Q 矩阵估计误差以提高参数估计精度，还应深入研究 Q 矩阵和项目参数的联合估计方法。特别地，已有研究大都聚焦于 DINA 模型和独立性属性结构。随着认知诊断模型的不断丰富，今后有必要探究其它认知诊断模型、不同属性层级结构、结合被试和项目特征等条件下的在线标定方法。

表 2 CAT 中项目参数在线标定方法

分类标准	方法	特点	适用情景
条件极大似然估计	MethodA、MethodB、FFMLE-A 和 ECSE-A	简单、易操作，需要大样本	传统 CAT/MCAT
	MLE-LBCI-A		传统 CAT
	CD-MethodA、MLE		CD-CAT
MMLE/EM 算法	OEM、MEM	计算复杂，耗时，不易收敛	传统 CAT 中二级和多级评分项目/MCAT
	CD-OEM、CD-MEM、MMLE		CD-CAT
贝叶斯算法	贝叶斯版本：方法 A，OEM 和 MEM	精度高、计算复杂，耗时	传统 CAT/MCAT
联合极大似然估计	JEA、SIE、SimIE、SIE-R、JEA-R、SIE-R-BIC、JEA-R-BIC RMSEA-N	联合估计 Q 矩阵和项目参数	CD-CAT

4 新题嵌入位置和终止规则

理论上，随着测验的进行，能力估计会越来越准确，将新题嵌入在测验的最后有利于提高参数估计精度。事实上，目前新题的嵌入设计有随机嵌入全卷不同位置（陈平，辛涛，2011a）、嵌入测验固定位置（Kingsbury, 2009）和嵌入全卷前部、中部、后部的随机位置（Zheng, 2014; He et al., 2019）几种方式。不同嵌入位置对参数标定产生不同的影响，目前还缺乏对不同方式进行系统的比较研究，也缺乏对新的嵌入方式的探索。例如，结合能力估计精度确定嵌入位置等等。

目前，在线标定中新题的终止规则主要有三种思路：基于作答新题的预设样本量规则（Ali & Chang, 2014; Zhu, 2006）、基于新题参数估计精度的规则（Ren et al., 2017）和基于参数估计稳定性规则（Kingsbury, 2009）。首先，虽有研究表明题目的样本量达到 500 就能提供比较准确的参数估计值，但样本量对新题标定的影响还有待深入研究。其次，应用新题参数估计精度的规则时还应设定考生样本上限，以避免某些题目一直不停止测验的风险。最后，参数估计稳定性规则容易受到参数估计方法的影响，需要考虑迭代不收敛的问题。新题何时停止施测决定参数标定的准确性，今后既应系统比较已有方法的表现，也应基于新的思路并结合多种信息深入研究新题的终止规则。

5 研究展望

项目增补对题库的开发和维护至关重要，在线标定技术的出现为项目增补开辟了新的途径。纵观国内外在线标定技术的研究，主要集中于在线标定设计和在线标定方法。本文首先从项目信息标准和考生样本约束的角度分类介绍在线标定设计、解析它们之间的关联和区别；然后，分别针对传统 CAT、MCAT 和 CD-CAT 介绍在线标定方法、分析相关发展趋势和思路。最后，对新题嵌入位置和终止规则的相关研究进行阐述并浅析。尽管在线标定设计和在线标定算法已取得丰富的研究成果，但仍有值得深思和改进的地方。总体上，未来还可以从以下方面开展深入研究。

5.1 基于信息量指标进一步探究在线标定设计

目前关于在线标定设计的研究主要围绕 D-优化展开。除了 D-优化设计，基于其它信息量指标都可以建构类似的关于项目参数的信息量指标，并用于选择新题。例如，未来可以将 KL 信息（kullback-leibler, KL; Xu et al., 2003）、后验加权 KL 信息（posterior weighted KL, PWKL; Cheng, 2009）、香农熵（shannon entropy, SHE; Wang & Chang, 2011）和互信息（mutual

information, MI; Mulder & van der Linden, 2009）等信息量转换到表征被试对项目参数提供的信息量，以此来构建在线标定设计的选题指标，都具有非常重要的价值。

5.2 深入项目增补方法的实证研究

目前关于项目增补的研究以理论为主，仅采用模拟实验检验在线标定方法以及在线标定设计的可行性和表现。尽管模拟实验可以控制实验条件开展重复实验，但很难保证与真实情境具备完全一致的测验条件。真实的测验情景中，利用在线标定技术标定新题参数，不仅可以验证已有的在线标技术的可行性，还能发现在模拟实验中难以发现的问题。因此，有必要在真实测验情景验证这些方法的表现。

5.3 深入研究 CD-CAT 在线标定方法

近二十年来，认知诊断模型得到了极大的丰富和发展，呈现出从单一测验条件到复杂测验条件模型，从低阶到高阶模型，从特殊到一般模型的发展特点。因此未来研究可以（1）基于一般化认知诊断模型，建构一般化的在线标定方法，并在多种特殊的诊断模型下比较它们的表现；（2）引入校正知识状态和项目参数估计误差的方法，改进已有在线标定方法；（3）进一步探究 Q 矩阵和项目参数的联合估计；（4）将已有的在线标定方法推广到多级评分项目、属性多级等复杂测验条件；（5）探索属性层级结构、模型复杂度、样本量和新题嵌入位置等因素对在线标定方法的影响。

5.4 加强 MCAT 中在线标定技术的研究

将传统 CAT 的在线标定设计和方法推广到 MCAT 中是未来研究的一种简单可行的方法。这种推广只是测量模型从单维 IRT 模型变化成多维 IRT 模型，被试的潜在特质由单维变成多维，公式推导的思想不变，具体计算发生相应变化。多维模型在实践中具有广泛应用，研究 MCAT 中项目参数的标定方法是今后研究的重要方向。

参考文献

陈平, 辛涛. (2011a). 认知诊断计算机化自适应测验中在线标定方法的开发. *心理学报, 43* (06), 710–724.

陈平, 辛涛. (2011b). 认知诊断计算机化自适应测验中的项目增补. *心理学报, 43* (07), 836–850.

陈平, 张佳慧, 辛涛. (2013). 在线标定技术在计算机化自适应测验中的应用. *心理科学进展, 21* (10), 1883–1892.

陈平. (2016). 两种新的计算机化自适应测验在线标定方法. *心理学报, 48* (9), 1184–1198.

谭青蓉. (2019). *CD-CAT 广义在线标定方法开发研究*(硕士学位论文). 江西师范大学, 南昌.

汪文义, 丁树良, 游晓锋. (2011). 计算机化自适应诊断测验中原始题的属性标定. *心理学报, 43*(08),

964–976.

熊建华, 罗慧, 王晓庆, 丁树良, (2018). 基于 GRM 的在线校准研究. *江西师范大学学报(自然科学版), 42*(01), 62-66.

游晓锋, 丁树良, 刘红云. (2010). 计算机化自适应测验中原始题项目参数的估计. *心理学报, 42*(7), 813–820.

Ali, U. S., & Chang, H. H. (2014). An item-driven adaptive design for calibrating pretest items, *ETS Research Report Series,* 1–12.

Ban, J. -C., Hanson, B. A., Wang, T. Y., Yi, Q. & Harris, D. J. (2001). A comparative study of on-line pretest item-calibration/scaling methods in computerized adaptive testing. *Journal of Educational Measurement, 38*(3), 191–212.

Berger, M. P. F. (1992). Sequential sampling designs for the two-parameter item response theory model. *Psychometrika, 57*(4), 521–538.

Berger, M. P. F. (1994). D-Optimal Sequential Sampling Designs for Item Response Theory Models.*Journal of Educational Statistics, 19*(1), 43–56.

Buyske, S. (1998). Optimal design for item calibration in computerized adaptive testing: The 2PL case. In N. Flournoy, et al. (Ed.), *New developments and applications in experimental design. Lecture Notes—Monograph Series, 34*. Haywood, CA: Institute of Mathematical Statistics.

Chang, Y. -C. I., & Lu, H. Y. (2010). Online calibration via variable length computerized adaptive testing. *Psychometrika, 75*(1), 140–157.

Chen, P. (2017). A comparative study of online item calibration methods in multidimensional computerized adaptive testing. *Journal of Educational and Behavioral Statistics ,42*(5), 559–590.

Chen, P., & Xin, T. (2013). Developing online calibration methods for multidimensional computerized adaptive testing. *Paper presented at the 78th Meeting of the Psychometric Society*, July, Arnhem, the Netherlands.

Chen, P., & Wang, C. (2015). A new online calibration method for multidimensional computerized adaptive testing, *Psychometrika, 81*(3), 674–701.

Chen, Y., Liu, J., & Ying, Z. (2015). Online item calibration for Q-matrix in CD-CAT. *Applied psychological measurement, 39*(1), 5–15.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika, 74*(4) 619–632.

Hassan, M. U., & Miller, F. (2019) Optimal item calibration for computerized achievement tests. *Psychometrika, 84*(4), 1101–1128 .

He, Y. H., Chen, P., Li, Y., & Zhang, S. M. (2017). A new online calibration method based on Lord's Bias-Correction. *Applied Psychological Measurement. 41*(6). 456-471.

He, Y. H., Chen, P., & Li, Y. (2019). New efficient and practicable adaptive designs for calibrating items online. *Applied Psychological Measurement, 44*(1), 3-16.

He, Y. H., & Chen, P. (2020). Optimal online calibration designs for item replenishment in adaptive testing. *Psychometrika*, *85*(1), 35-55.

Jones, D. H., & Jin, Z. Y. (1994). Optimal sequential designsfor on-line item estimation. *Psychometrika, 59*(1), 59–75.

Kang, H. A., Zheng, Y., & Chang, H. H. (2020). Online calibration of a joint model of item responses and response times in computerized adaptive testing.*Journal of Educational and Behavioral Statistics, 45*(2),175-208.

Kingsbury, G. G. (2009). Adaptive item calibration: A process for estimating item parameters within a computerized adaptive test. In D. J. Weiss (Ed.), P*roceedings of the 2009 GMAC conference on computerized adaptive testing*(pp.1–15). Retrieved from http://www.psych.umn.edu/psylabs/CATCentral/

Makransky, G. (2009). An automatic online calibration design in adaptive testing. *Paper presented at the 2007 GMAC Conference on Computerized Adaptive Testing*, McLean, USA.

Mulder, J., & van der Linden, W. J. (2009, June). Multidimensional adaptive testing with optimal design criteria for Item Selection. *Psychometrika, 74* (2), 273–296.

Ren, H., van der Linden, W. J., & Diao, Q. (2017). Continuous online item calibration: Parameter recovery and item utilization. *Psychometrika, 82*(2), 498–522.

Stefanski, L. A., & Carroll, R. J. (1985). Covariate measurement error in logistic regression. *Annals of Statistics, 13*(4), 1335–1351.

Stocking, M. L. (1988). Scale drift in on-line calibration (Research Rep. 88-28). Princeton, NJ: ETS.

van der Linden, W. J., & Ren, H. (2015). Optimal bayesian adaptive design for test-item calibration. *Psychometrika, 80*(2), 263–288.

Wainer, H., & Mislevy, R. J. (1990). Item response theory, item calibration, and proficiency estimation. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (Chap. 4,pp. 65–102). Hillsdale, NJ: Erlbaum.

Wang, C., & Chang, H. H. (2011). Item selection in multidimensional computerized adaptive testing–gaining information from different angles. *Psychometrika,76*(3), 363–384.

Xiong, J., Ding, S., Luo, F., & Luo, Z. (2020). Online calibration of polytomous items under the graded response model. *Frontiers in psychology, 10*(1), 3085.

Xu, X. L., Chang, H. H., & Douglas, J. (2003). *A simulation study to compare CAT strategies for cognitive diagnosis*. Paper presented at the annual meeting of National Council on Measurement in Education, Montreal, Canada.

Zheng, Y. (2014). *New methods of online calibration for item bank replenishment* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign, Champaign, IL.

Zheng, Y. (2016). Online calibration of polytomous items under the generalized partial credit model. *Applied Psychological Measurement, 40*(6), 434-450.

Zheng, Y., & Chang, H. H. (2017). A comparison of five methods for pretest item selection in online calibration. *International Journal of Quantitative Research in Education, 4*(1), 133–158.

Zhu, R. (2006). *Implementation of optimal design for item calibration in computerized adaptive testing (CAT)* (Unpublished doctoral dissertation). University of Illinois, Urbana.

Online calibration based on computerized adaptive testing: Design and method

ZHANG Xueqin[1]; MAO Xiuzhen[1]; LI Jia[1]

(Institute of Educational Science, Sichuan Normal University, Sichuan Chengdu, 610066)

Abstract: Item replenishment is essential for item bank development and maintenance, where new items' parameter calibration plays a significant role. Two core techniques of item replenishment under the circumstances of computerized adaptive testing (CAT) are: 1) online calibration design; 2) online calibration method. The former investigates the administration way of new items, while the later explores parameter estimation methods. This paper aims to clarify the development ideas and contexts of online calibration design and online calibration method. Additionally, their characteristics, relations and performance were illustrated and evaluated in details. At the end, several future research directions were pointed out. It is important to further study online calibration design based on different information indicators and online calibration methods based on joint estimations and error corrections. Moreover, future study could explore the online calibration technique in cognitive diagnostic CAT(CD-CAT) and multidimensional

CAT(MCAT), as well as the empirical applications of item replenishment.

Key words: computerized adaptive testing; cognitive diagnostic theory; item replenishment; online calibration design; online calibration method